Scaffolding and Running Your First Scrapy Web Crawler Project

Eduardo Freitas

BUSINESS AUTOMATION & DATA CAPTURE SPECIALIST

Overview

Introduction to Scrapy

Architecture

Beautiful Soup

Creating and scaffolding a Scrapy project

Introduction to Scrapy

Scrapy is a framework,
not a library.

Framework vs. Library

Framework	Library
A framework defines how your application works	Your application invokes the library, it is not defined by it
A framework invokes your code	Your application invokes the library
Your application can only use one framework at a time	Your application can use multiple libraries at a time

Scrapy Pillars

Asynchronous

Control

Scrapy Advantages

- Fault tolerant
- Parallel execution
- Reliable
- Control
- Speed

Scrapy Architecture

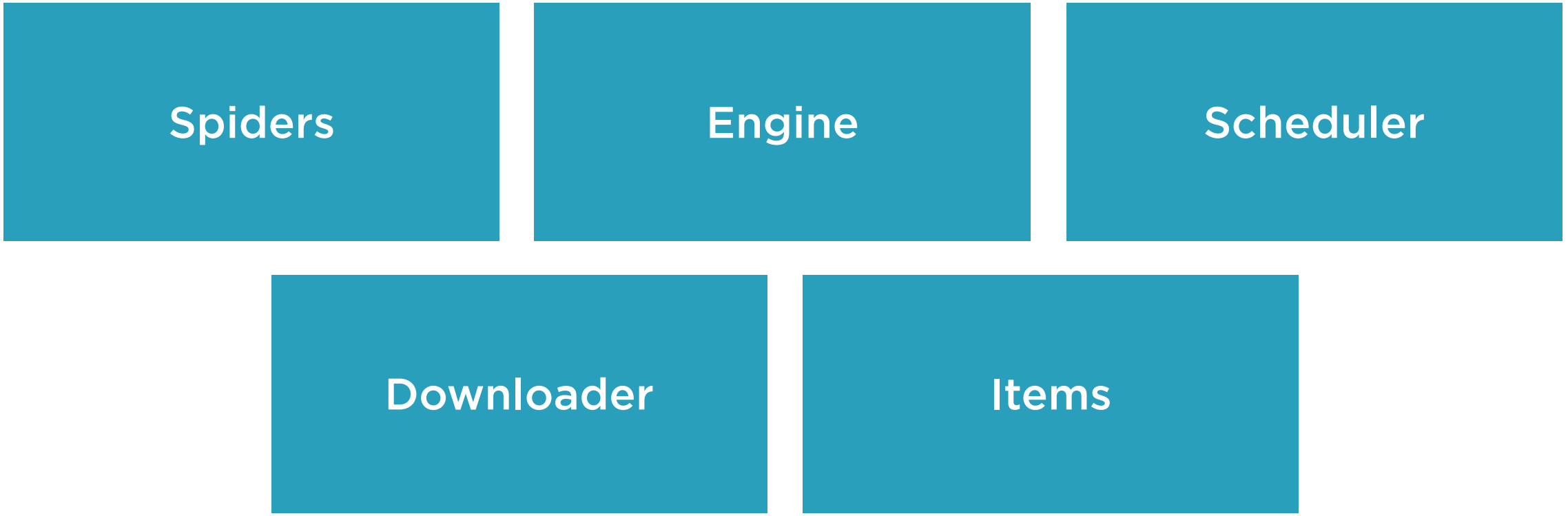
Fined-grained Control

Limit download requests

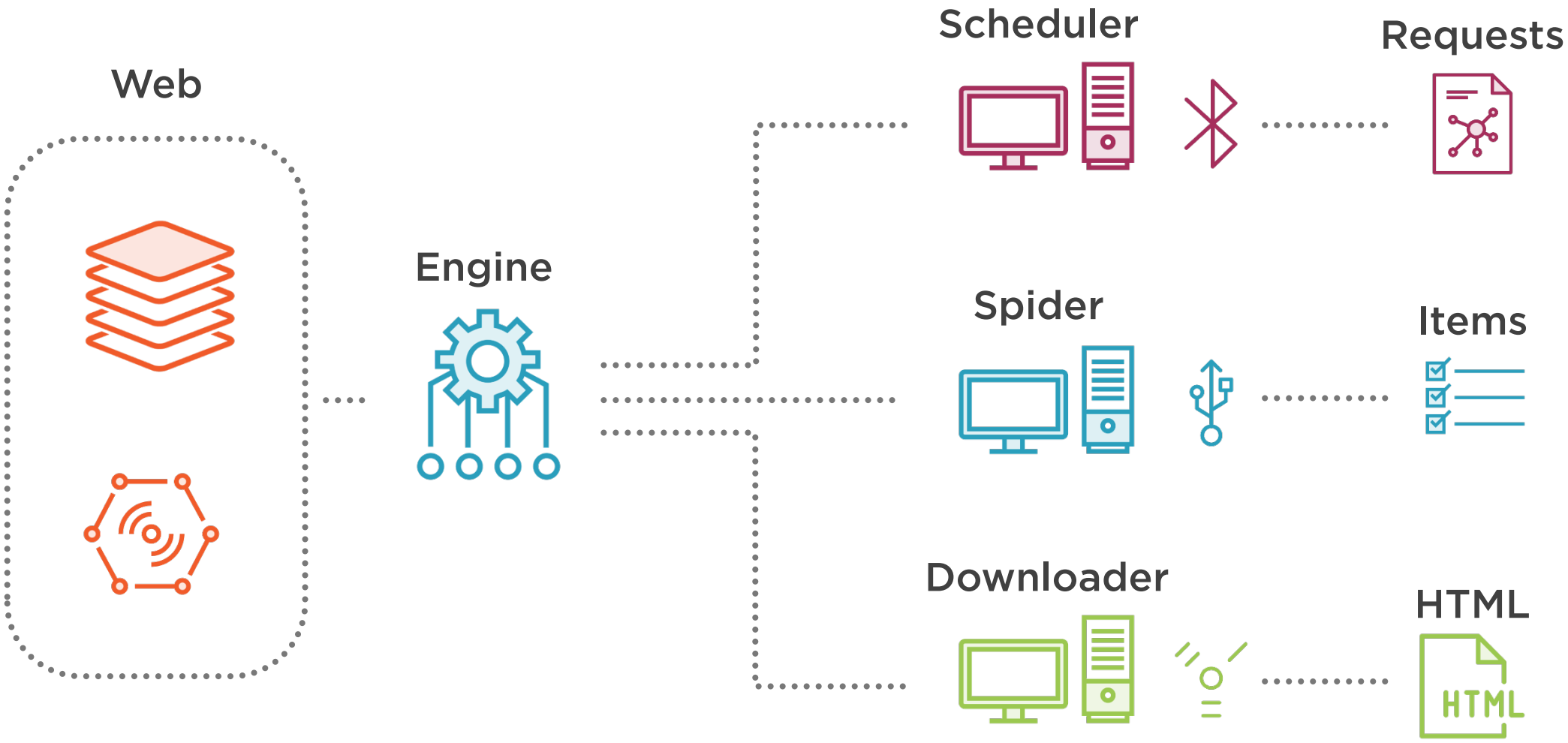
Concurrent connections

Automatic throttling

Scrapy Architecture



Scrapy Architecture



Beautiful Soup

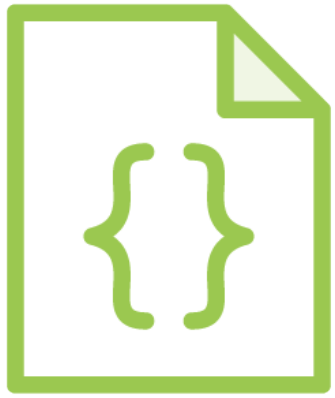
Python package for parsing for understanding and parsing HTML and XML, including incorrect markup and missing tags.

Beautiful Soup

Full HTML

Solid parser

Advantages



- Simple and intuitive to use
- Able to parse the entire HTML tree
- Overcomes incorrect markup
- Completes missing tags
- Stronger than regular expressions
- Uses popular parsers

Objects

Beautiful Soup has four different kind of objects

Tag – indicates an XML or HTML tag in the original document

NavigableString – specifies a bit of text within a tag

BeautifulSoup - represents the parsed document

Comment - special type of NavigableString

Beautiful Soup

Official Documentation

https://www.crummy.com/software/BeautifulSoup/

Demo

Creating and scaffolding a new Scrapy project

Summary

- Framework vs. library
- Advantages of Scrapy
- How Scrapy is architected
- Beautiful Soup overview
- Beautiful Soup objects
- Scaffolded a new project